A Ragozási szótártól a NooJ morfológiai moduljáig

Vajda Peyter, Nagy Viktor, Dancsecs Erzsébet

MTA Nyelvtudományi Intézet, Budapest VI. Benczúr u. 33. Pf. 701/518 H-1399
{vajda,nagyv,mano}@nytud.hu

Kivonat A cikk a *NooJ* számítógépes nyelvészeti fejlesztőkörnyezet magyar morfológiai moduljának fejlesztését mutatja be. A morfológiai elemző alapja Elekfi László Magyar ragozási szótárának a Nyelvtudományi Intézetben megvalósított számítógépes változata. A morfológia leírásában a paradigmatikus ábrázolás helyett át kellett térnünk a jegy alapú osztályozásra. Az implementálásban a *NooJ* véges állapotú technikáját használtuk.

1. Bevezetés

A Nyelvtudományi Intézet Korpusznyelvészeti osztálya azt a célt tűzte ki, hogy megvalósítja az *Intex*, illetve utódja, a *NooJ* nyelvészeti fejlesztőkörnyezet (továbbiakban: *Intex/NooJ*) alá a magyar nyelvi modult, vagyis azokat az alapnyelvtanokat, amelyekre majd építhet a rendszer leendő felhasználói közössége. A dolgozat által bemutatott morfológiai elemző és lemmatizáló ennek a modulnak lesz a része.

2. Az Elekfi-rendszer

Az általunk kidolgozott morfológiai elemző alapját Elekfi László *Magyar ragozási szótára* ([2]), illetve az egyelőre csak kéziratban létező, a ragozási szótárnál jóval részletesebb *Szókincsünk nyelvtani alakrendszere* ([3]) című munkája adta. Az anyag gépre vitelével morfológiai adatbázis épült.A ragozási szótár címszóanyaga megegyezik a *Magyar Értelmező Kéziszótár* első kiadásának címszóanyagával. A morfológiai adatbázis tőtára a lexémák tőallomorfjait tartalmazza.

Az Elekfi-féle rendszer a lexémákat paradigmaosztályokba sorolja. Az osztályok kétdimenziós elrendezést mutatnak. Az egyik dimenziót az elölségi, illetve a kerekségi harmónia adja (betűjelek az osztálykódokban: A: „mély" (hátulképzett), B: „magas" (elölképzett), kerekítetlen, C: kerekített; igéknél kisbetűvel jelölve), a másik dimenziót a lexéma tövének egyéb komplex tulajdonságai adják. A második dimenzió az igéknél tizenkilenc, a névszóknál harminchat csoportot határoz meg. Ezek a csoportok finom paradigmatikus különbségek szerint további alcsoportokra oszlanak.

Ez a fajta paradigmaosztályozás több hátránnyal bír. A paradigmák közötti számos megegyezés és rendszerszerű különbség rejtve marad, hiszen a két dimenzió csak két tulajdonságot képes ábrázolni (bővebben lásd [5]). A paradigmák közötti finom különbségek pedig megnövelik az alcsoportok számát, hiszen minden egyedi ragozási sor egyedi osztályt kell, hogy alkosson. A teljes rendszer több mint 1700 paradigmaosztályt tartalmaz.

Ilyen sok osztály karbantartása és implementációja nehézkes. Egy megfelelő véges állapotú eszköz képes lehet optimális, redukált automata előállítására, azonban mind az *Intex*, mind a *NooJ* csak korlátozottan rendelkezik ilyesfajta képességekkel. A rendszer inflexiós modulja a redukálást generálással oldja meg: a szabályrendszer alapján előállítja az összes szótári szó lehetséges alakjait, melyeket automatába tömörít. Ez a megoldás megfelelő a szegényes inflexiót felmutató nyugat-európai nyelveknek, a magyarnak semmiképp. Még ha mellőztük is a rekurziót a magyar morfológiából, a néhány tízezres szótár több tízmillió lehetséges szóalakjának tömörítése erőforráskorlátokba ütközött.

Az Elekfi-rendszer további korlátja, hogy szűk a számba vett toldalékok köre. A paradigmák az inflexiós toldalékok mellett csupán néhány produktív képzőt tartalmaznak (igenévképzők, *-hAt*, *-Ás*, műveltető, *-(j)Ű*). További képzők hozzáadása a meglévővel ortogonális osztályozást igényel, hiszen gyakran olyan szemantikai tulajdonságoktól függ, hogy egy képző hozzájárulhat-e egy adott tőhöz, amelyekhez nem rendelhetők Elekfi-paradigmák.

3. A morfológiai leírás módszere

Mint az előbbiekben láttuk, a paradigmatáblák nehézkesen kezelhetők, nagyfokú redundanciát tartalmaznak. A kétdimenziós osztályozás miatt több paradigma tartalmazhat hasonlóan viselkedő elemeket, de az ilyen hasonlóságok csak akkor ragadhatók meg, ha ezen elemeket bármilyen szempont szerint be tudjuk sorolni egy-egy csoportba.

Ennek elérése érdekében jegyeket alakítottunk ki, melyek segítségével egymástól független szempontok szerint tudtuk a szavakat osztályozni. Majd felépítettünk egy, a fenti jegyeket használó véges állapotú transzducert. Ezzel a paradigmaosztályok fent említett hátrányait kiküszöböltük. Az alábbiakban részletesebben foglalkozunk az *Intex/NooJ* rendszer eszközkészletével (bővebben lásd [6]), valamint a transzducer felépítésével, példákon keresztül megmutatva a felmerült nehézségek megoldását.

3.1. Az Intex végesállapotú technológiája

Az *Intex* többek között szótárak, a derivációs és inflexiós morfológia, és a szintaxis leírásához, valamint korpuszok feldolgozásához nyújt eszközöket.

Az *Intex* a feldolgozás valamely pontján minden erőforrást (szótárak, nyelvtanok, stb.) végesállapotú transzducerrel (*Finite State Transducer, FST*) reprezentál. Egyszerűbb végesállapotú automaták leírására alkalmazhatunk reguláris kifejezéseket, bonyolultabb nyelvtanok beviteléhez pedig felhasználói felületként

egy gráf-építő modult használhatunk. Egy gráf egy FST-t reprezentál, amely esetünkben a (jólformált) bemeneti szöveghez morfológiai információt társít. Egy transzducer azonban nemcsak úgy fogható fel, hogy egy karaktersorozathoz egy másikat rendel, hanem úgy is, mint egy karaktersorozatot elfogadó véges állapotú automata (bővebben lásd [1]). Ez megfelel a morfológiai elemző feladatának, mely egyszerre elfogad és elemez egy szóalakot.

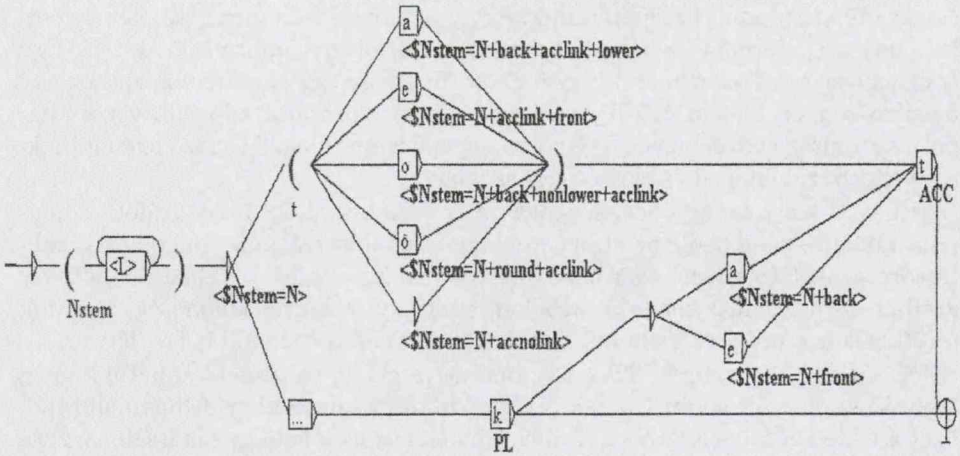
Ellentétben a transzducerek szokásos gráf alakjával, az Intex-gráfok csomópontjaikban – nem pedig az átmenetekben – tartalmazzák a felismerendő karaktersorozatokat (betűket) és a hozzájuk tartozó kimenetet, továbbá állapotok a gráfban nem jelennek meg. Ez azonban csak egy technikai különbség: a gráfok fordításakor a program minimalizált determinisztikus automatát hoz létre.

Az *Intex* kiterjesztett FST-ket is használ: ezekben változókat adhatunk meg, melyek az elemzés során kapnak értéket, majd a kimenetben felhasználhatjuk őket a bemenet módosításához. Ennek többek között a kétjegyű mássalhangzóra végződő szavaknál van szerepe, ahol egyes toldalékok esetében a tő és a toldalék konkatenációjával nem áll elő a helyes alak (pl.: *lány+nyá* ≠ *lánnyá*), ezért a bemenetből egy karaktert törölni kell.

Jegyalapú morfológia. A jegyalapú morfológiák sajátossága, hogy egy-egy (relatív) tő és (relatív) toldalék jegyeinek unifikációjával, a jegyek ellenőrzésével áll elő a toldalékolt alak. Az *Intex/NooJ* eszközeivel azonban csak a szótő és az első toldalék jegyeit lehet közvetlenül egyeztetni, az egymást követő toldalékokét már nem, mivel a toldalékok a szótárban nem szerepelhetnek, így ott jegyeket sem lehet hozzájuk rendelni. A toldalékokhoz rendelt megszorításokat pedig a szótárban kell ellenőrizni (lásd a 3.1. pontot). Ez az oka annak, hogy azokban az esetekben, ahol egy toldalékot relatív tőként egy másik toldalék követ, a gráfban új csomópontokat kell felvennünk a kapcsolódó toldalékhoz.

Ez a jelenség megfigyelhető például a többesszám és a tárgyeset kapcsolódásánál, amit az 1. ábra illusztrál. A tárgyeset viselkedésének leírását a gráfunkban két esetre bontottuk. Az első eset az, amelyben a tárgyeset közvetlenül az abszolút szótőhöz kapcsolódik. Ekkor a tárgyrag a *-(V)t* alakot veheti fel, ez a gráfban öt útvonalat jelent, amelyeket a megfelelő kötőhangokat kiválasztó jegyek engedélyeznek. A második esetben, nyitótőként viselkedő toldalékok után a tárgyeset alakja viszont már *-At*. Az *Intex/Nooj* eszközeivel nem tudunk olyan feltételt megfogalmazni, mely megadná, hogy a többesszám kötőhangjától függően miyen kötőhangot vár a tárgyeset.

Lexikai megszorítások. Az *Intex/NooJ* rendszerben lexikai megszorítások adhatók meg a szóalakok különböző részeire. A megszorítások lexikai jegyek formájában jelennek meg, és a szótár segítségével kerülnek ellenőrzésre. A lexikai megszorításokban hivatkozhatunk az elemzés során előzőleg definiált változók értékeire. Ennek segítségével oldottuk meg a szótő és a toldalékok elválasztását (lásd az 1. ábrát). A gráfban egy hurok inkrementálisan egy `Nstem` változóhoz rendeli a bemeneti sztring betűit, majd, hogy mi kerül végül szótőként ebbe a



1. ábra. A tárgyeset kezelése

változóba, azt a `<$Nstem=N>` megszorítás dönti el, amely a szótő meglétét ellenőrzi a szótárban (természetesen ezután az automatának még el kell jutnia a végállapotba).

Továbbá, ha a transzducer egy csomópontjában az `ö/<stem=N+round>` információ szerepel, akkor a gráfnak ezen az ágán csak azok a szavak tudnak továbblépni, amelyek a feldolgozás ezen pontján egy *ö* betűt tartalmaznak, és a szótárban a `+round` jeggyel szerepelnek. Az, hogy mind az *Intex*ben, mind a *NooJ*-ban csak egy jegy meglétét lehet ellenőrizni, annak hiányát nem, azt jelenti, hogy csak egyértékű jegyek adhatók meg. Ezért nem a (morfofonológiában) megszokott kétértékű jegyeket használtuk, hanem jegy-párokat (pl.: `+round`, `+unround`)

.

3.2. A jegyek kialakítása

Az általunk használt jegyek kialakításánál [5]-re és az Elekfi-rendszerre támaszkodtunk. Természetesen felhasználtuk a rendszer egyik dimenzióját adó, a magánhangzó-harmónia elemzéséhez szükséges jegyeket. Ezen túlmenően a kétdimenziós csoportok másik dimenzióját adó összetett tulajdonságokat (pl.: „zárt kötőhangzós főnevek puszta tárgyraggal"), valamint e csoportokon belül található rendszeres különbségtételeket (pl.: a *j* hang megjelenése a névszók harmadik személyű birtokos alakjánál) is feldolgoztuk. Ezek a tulajdonságok okozzák az Elekfi-rendszer redundanciáját és bonyolultságát, ezért ezekből egymástól független jegyeket alakítottunk ki, amelyek már alkalmasak egy többdimenziós osztályozás felállítására.

A tőallomorfia kezelése. A létrejött jegyek közül külön említést érdemelnek a tőallomorfiára vonatkozóak. Bár egy transzducer elvben képes szimbólumokat

törölni vagy beszúrni az elemzendő karaktersorozatba, ily módon kezelve a tőallomorfiát, ezt a kérdést az Elekfi-rendszerhez hasonlóan tőallomorfok használatával oldottuk meg. A morfológiai adatbázisban lévő töveket az ismert morfofonológiai osztályokba (pl.: hangkivető, rövidülő, bővebben lásd [4]) soroltuk, és a szótárban erre utaló jegyeket társítottunk hozzájuk. A szótárba a lexémák mellett a kötött tőallomorfok is bekerültek az erre utaló jegyekkel. Ezután megvizsgáltuk, hogy az allomorfok hogyan viszonyulnak az egyes toldalékokhoz. A különböző típusú tőváltozat-osztályokat tovább csoportosítottuk az alapján, hogy mely toldalékokhoz járulnak a szótári alakjukal (BASE), melyekhez pedig allomorfjukkal (OBL). A névszói tövek csoportosításának egy részlete látható az 1. táblázatban.

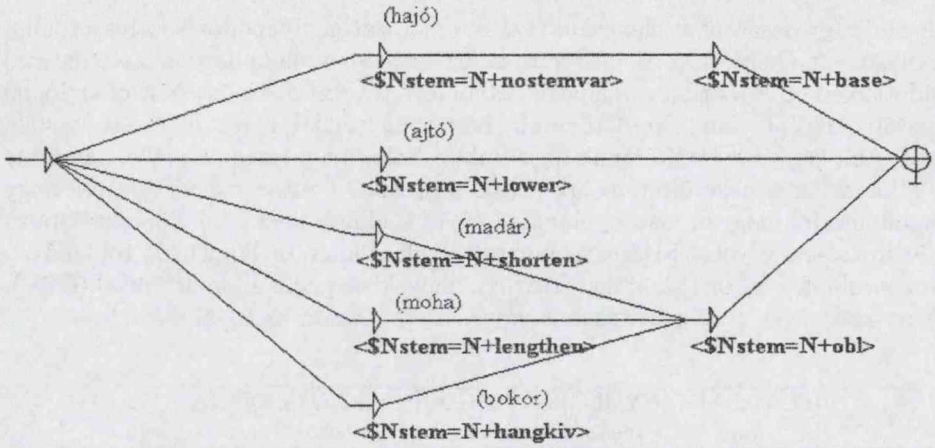
	EGYALAKÚ *hajó*	NYÚLÓ *moha*	RÖVIDÜLŐ *madár*	HANGKIVETŐ *bokor*	HANGVÁLTÓ *ajtó*
PS[et][12], ACC, PL.NOM, DIS/SOC,	BASE	OBL	OBL	OBL	BASE
PS[et]3	BASE	OBL	OBL	OBL	OBL/BASE
SUP	BASE	OBL	BASE	OBL	BASE
esetek, POS	BASE	OBL	BASE	BASE	BASE
NOM	BASE	BASE	BASE	BASE	BASE

1. táblázat. Tőosztályok

A táblázatból leolvasható, hogy ezen öt névszói tőosztály esetében öt csoportba kell beosztanunk a toldalékokat. A megvalósítás szintjén ez annyit jelent, hogy minden egyes csoporthoz egy-egy beágyazott gráfot készítettünk, mely az adott toldalék(ok)hoz kapcsolódik, és az összes tőtípus erre vonatkozó viselkedését írja le – azt, hogy a szótári alakkal, vagy pedig a kötött tővel jár-e együtt. Például a harmadik személyű birtokos alak algráfjában (lásd a 2. ábrán) a nem változó tőtípusok a szótári alakot (`+base`) kapják, a hangváltó tövek esetében mindkét allomorf lehetséges (*ajtója*, *ajtaja*, ill. *ajtói*, *ajtai*), ezért ide nem kerül megszorítás, a többi tőtípus pedig a nem szótári alakkal (`+obl`) jár együtt.

Természetesen további tőosztályok, rendhagyó paradigmák felvételével a toldalékok csoportosítása is megváltozik.

Rendhagyó ragozású tövek. Az Elekfi-rendszer 9-es csoportjai a rendhagyó szavakat tartalmazzák, amelyek az általános paradigmákba nem illeszthetőek be. Ezért ragozásuk leírása több nehézségbe is ütközik. Egyrészt a szótárban a szabályos paradigmákhoz kialakított jegyek alkalmazása nem elégséges, új jegyek felvétele viszont túlságosan bonyolulttá tenné a szótár és a főgráf szerkezetét. Megvizsgálva a csoportba tartozó főneveket, három lehetőség állt elő. Készíthetünk algráfokat, megpróbálhatjuk beilleszteni a töveket szabályos paradigmákba, vagy statisztikai okok miatt figyelmen kívül is hagyhatjuk őket.



2. ábra. A tőallomorfia kezelése

Általában véve ezeknek a töveknek a viselkedését célravezető algráfokban leírni, majd beilleszteni őket a főgráf szerkezetébe. Ebben az esetben csak azt a jegyet kell kiosztanunk, amely a rendhagyó paradigmát meghatározza, ugyanakkor alkalmazhatjuk a már meglévő jegyeket is pl. a hangrendre vonatkozóan. A tő viselkedése szempontjából például azonos, de hangrendileg különböző a 9A3 (*bátya*) és 9B3 (*néne*).

Helyenként alkalmazható eljárás a tövek már meglévő paradigmákba történő besorolása is. Például az egyetlen csoportba sorolt *fiú* főnév esetében szétválasztottuk belőle a *fi* és *fiú* töveket, amelyek már beilleszthetők egy-egy szabályos paradigmába. Vagy például a *h*-ra végződő tövek szintén beleillenek egy-egy szabályos paradigmába, viszont két esetragnál opcionálisan más paradigma toldalékait is megkaphatják (*cseh-vel*, de **cseh-hel*, ill. *méh-hel*, *méh-vel*), ezért ez a tőallomorfiánál látott módszerrel oldható meg.

A 9-es csoport nem minden alcsoportja került besorolásra hatékonysági szempontok miatt. Megvizsgáltuk, mely paradigmákhoz hány szó tartozik, illetve azt, hogy az alapszó milyen előfordulást mutat az MNSZ-ben. Azokat a paradigmákat hagytuk el, amelyek alá egyetlen szó tartozik, és a Magyar Nemzeti Szövegtárban nem fordul elő (ilyen például a *moholy* szó). Ezekre a perifériikusnak minősíthető tövekre nem készült külön algráf, illetve nem kerültek besorolásra. Hasonlóan jártunk el a régiesnek tekinthető szavakkal is (pl.: *tereh*).

Figyelembe vettük viszont azokat a paradigmákat, amelyekbe ugyan csak egy vagy két szó tartozik, viszont több előfordulást találtunk rá – ilyen pl. a *kehely*, mely csak alanyesetben 86-szor fordul elő az MNSZ-ben -, vagy intuitíve fontos szónak tartottuk ragozott alakjai miatt (mint pl. a *bátya* szó, mely ugyan csak 39-szer található meg az MNSZ-ben, viszont birtokos személyjeles alakjai jóval gyakoribbak, mert pl. a *bátyja* alak is ebből a paradigmából áll elő). Bár a

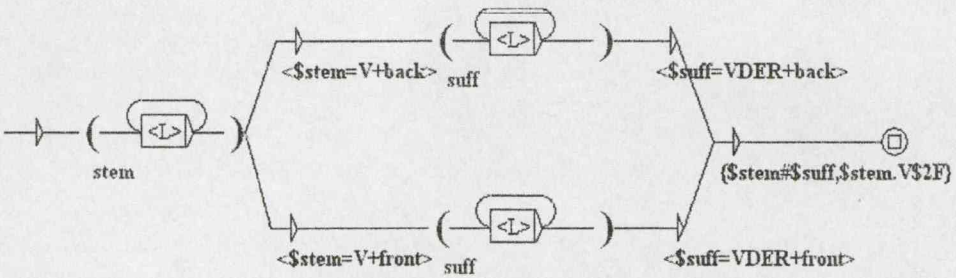
rendhagyó igék kimerítő vizsgálatára még nem került sor, valószínűsíthető, hogy a főneveknél alkalmazott eljárásokat itt is hasznosíthatjuk majd.

3.3. A szótár

A morfológiai elemző modul szótárát az Elekfi-rendszer szótárából alakítottuk ki. Az eredeti szótárban a számunkra lényeges információ a szótövek mellet a szófaj, az esetleges tőallomorfok és az illető szónak az Elekfi-rendszerbeli paradigmakódja volt. Az általunk használt szótárban a szótövek és a tőallomorfok külön bejegyzésként szerepelnek a szófaj és a szóra vonatkozó jegyek feltüntetésével. A paradigmakódok közvetlen jegyekre alakítása csak nehezen, kézzel lett volna elvégezhető, ezért más módszert választottunk. Mivel az Elekfi-adatbázisból a rendszer bármely szavának bármely ragozott alakja kinyerhető, ezért minden olyan alakot meg tudtunk vizsgálni, amely alapján eldönthető, hogy az egyes jegyek szempontjából, hogy viselkedik az adott szó. Például annak eldöntéséhez, hogy egy szó megkapja-e a `+acclink` jegyet (ami azt jelzi, hogy a tárgyesetben van-e kötőhangja), elegendő megvizsgálni a tárgyesetű alakját. Ily módon automatikusan az összes paradigmához hozzá tudtuk rendelni a viselkedését meghatározó jegyhalmazt.

4. Képzés

Fentebb láttuk, hogy a nemterminális toldalékoknak a tőlük jobbra álló morfémával szemben támasztott jegyigényeit nem tudjuk megfogalmazni, helyette új csomópontokat és utakat kell felvenni a gráfban. Emiatt a képzők felvételével sokszorosára növekedne az elemzőgráf bonyolultsága. Ehelyett azt tervezzük, hogy a képzőket fiktív tőként vesszük fel a szótárba. A 3. ábrán látható gráf felel a szótő és a képző összekapcsolásáért, ha a képzőnek csak egy magas és egy mély allomorfja van. A *suff* változóba kerülő részsztring már a képző továbbtoldalékolt alakja.



3. ábra. A képzés

5. Összefoglalás

A dolgozatban felvázoltunk egy fejlesztés alatt álló morfológiai elemzőt *Intex/NooJ* alá, amelyet terveink szerint hozzáférhetővé teszünk bárki számára nonprofit, kutatási célokra. A fejlesztőeszköz nem ideális a feladatra, hiszen elsősorban nyugat-európai nyelvekre készült. A magyar modul létrehozásának szándéka azonban ösztönzőleg hat az *Intex/NooJ* fejlesztőire, akik igyekeznek bővíteni a rendszer szolgáltatásait, hogy alkalmas legyen a magyar morfológia implementálására is.

A fejlesztés során szakítanunk kellett az Elekfi-rendszerrel, és jegyalapú osztályozást kellett alkalmaznunk. Ezáltal rugalmasabb, könnyebben karbantartható adatbázist kaptunk, amely alkalmas lehet más eszközökkel történő morfológiai elemző létrehozására is.

Hivatkozások

1. Dienes Péter: A fonológia modellezése végesállapotú eszközökkel. In: Lexikalista elméletek a nyelvészetben. Szerk.: Kálmán László, Trón Viktor, Varasdi Károly. Tinta Könyvkiadó, Budapest 2002. 17-51.
2. Elekfi László: Magyar ragozási szótár. MTA Nyelvtudományi Intézete, Budapest 1994.
3. Elekfi László: Szókincsünk nyelvtani alakrendszere. Kézirat. 1986.
4. Nádasdy Ádám és Siptár Péter: A magánhangzók. In: Strukturális magyar nyelvtan 1. Fonológia. Szerk.: Kiefer Ferenc. Akadémiai Kiadó, Budapest 1994. 42–182.
5. Prószéky Gábor: A magyar morfológia számítógépes kezelése. In: Strukturális magyar nyelvtan 3. Morfológia. Szerk.: Kiefer Ferenc. Akadémiai Kiadó, Budapest 2000. 1021–1063.
6. Silberztein, Max: The Intex Manual. `http://intex.univ-fcomte.fr/downloads/Manual.pdf`